Heart Disease Prediction Using Machine learning(Random Forest Algorithm and XGBOOST Algorithm

NAME: GOLLAPUDI SHESHA GAYATHRI

COLLEGE / DEPARTMENT: GRAPHIC ERA HILL UNIVERSITY / CSE

DATE: 25, 2025

Introduction

- Heart disease is one of the leading causes of death worldwide.
- Early detection can significantly reduce mortality and improve treatment.
- Traditional diagnosis depends on doctor's experience and manual analysis, which can be time-consuming and error-prone.
- Machine learning can:
  - Automate prediction
  - Handle large and complex datasets
  - Discover hidden patterns in medical data
- Objective:
  - Predict presence of heart disease using patient clinical data
  - Compare Random Forest and XGBoost models

Problem Statement

- Problem Statement
- Can we predict whether a patient has heart disease based on clinical attributes?
- Input features:
  - Age, Sex, Chest pain type, Blood pressure, Cholesterol, Fasting blood sugar, ECG results, Max heart rate, Exercise angina, ST depression, Slope of ST, Number of vessels, Thallium test
- Output (target):
  - Heart Disease:
    - 1 → Presence of heart disease
    - 0 → Absence of heart disease
- Goal:
  - Build and compare two ML models (Random Forest and XGBoost) for accurate prediction

Dataset Description

- Dataset Description
- Dataset: Heart Disease Prediction dataset
- Total records: [Insert number, e.g., 303]
- Features: 13 input variables + 1 target
- Key Features:
- Age
- Sex (0/1)
- Chest pain type
- Resting blood pressure (BP)
- Serum cholesterol
- Fasting blood sugar > 120 mg/dl
- Resting ECG results

- Maximum heart rate
- Exercise induced angina
- ST depression
- Slope of ST segment
- Number of major vessels colored by fluoroscopy
- Thallium stress test result
- Target Variable:
- Heart Disease: Presence (1) / Absence (0)

Data Preprocessing

- Data Preprocessing
- Steps performed:
- Load Data
  - Read CSV file using `pandas`
  - Assign meaningful column names
- Target Encoding
  - Map target values:
    - `'Presence'` → `1`
    - `'Absence'` → `0`
- Train-Test Split
  - Split data into training and testing sets
  - Ratio: 80% training, 20% testing
  - Stratified split to maintain class distribution
- Feature Selection
  - Use all 13 clinical features as input
  - No missing values (or handled if any)

Machine Learning Algorithms Used

Two ensemble algorithms:

Random Forest Algorithm

XGBoost Algorithm

- Random Forest
  - Ensemble of multiple decision trees
  - Each tree is trained on a random subset of data and features
  - Final prediction by majority voting
  - Advantages:
    - High accuracy
    - Handles non-linear relationships
    - Reduces overfitting
    - Works well with medical datasets

- XGBoost (Extreme Gradient Boosting)
  - Advanced boosting algorithm
  - Builds models sequentially, each correcting errors of the previous one
  - Uses gradient descent to minimize loss
  - Advantages:
    - Very high prediction accuracy
    - Efficient and fast
    - Handles missing values well
    - Widely used in real-world applications

Random Forest – How It Works

- Random Forest – How It Works
- Ensemble of Decision Trees
  - Many decision trees are created
  - Each tree uses a random subset of data and features
- Prediction Process
  - Each tree predicts the class (0 or 1)
  - Final output is decided by majority voting
- Why It Works Well
  - Reduces variance and overfitting
  - Robust to noise and outliers
  - Good for structured medical data

XGBoost – How It Works

- Boosting Approach
  - Models are built sequentially
  - Each new model focuses on correcting errors of the previous model
- Gradient Descent
  - Uses gradients to minimize prediction error
  - Learns from mistakes to improve accuracy
- Why It Works Well
  - Very high accuracy on structured data
  - Efficient and scalable
  - Handles missing values and complex patterns
  - Popular in competitions and real-world systems

Model Training and Evaluation

- Training Setup
- Both models trained on the same training set
- Hyperparameters used:
  - Random Forest: `n_estimators=100`, `random_state=42`
  - XGBoost: `use_label_encoder=False`, `eval_metric='logloss'`, `random_state=42`
  - Evaluation Metrics
- Accuracy
- Precision
- Recall
- F1-score
- Confusion Matrix
- Evaluation on:
- Test dataset (20% of data)

Results – Random Forest

- Results – Random Forest
- Random Forest gives stable and reliable results.
- Accuracy: [Insert value, e.g., 85%]
- Classification Report:

Class	Precision	Recall	F1-score
Absence (0)	[P0]	[R0]	[F10]
Presence (1)	[P1]	[R1]	[F11]

- Confusion Matrix:

	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

Results - XGBOOST

- Accuracy: [Insert value, e.g., 88%]
- XGBoost achieves higher accuracy and captures complex patterns in heart disease data.
- Classification Report:

Class	Precision	RECALL	F1-score
Absence (0)	[P0]	[R0]	[F10]
Presence (1)	[P1]	[R1]	[F11]

- Confusion Matrix:

	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

Model Comparison

- Model comparison

Metric	Random Forest	XGBoost
Accuracy	[RF Acc]	[XGB Acc]
Precision (1)	[RF P1]	[XGB P1]
Recall (1)	[RF R1]	[XGB R1]
F1-score (1)	[RF F11]	[XDB F11]

- Conclusion:
- XGBoost outperforms Random Forest in terms of accuracy and prediction capability.
- XGBoost is better at capturing complex patterns in the heart disease dataset.

Conclusion

- Machine learning can effectively predict heart disease using patient clinical data.
- Among the two models:
  - Random Forest provides stable and reliable predictions.
  - XGBoost achieves higher accuracy and better captures complex patterns.
- These models can:
  - Assist doctors in early diagnosis
  - Support decision-making in clinical settings
- This project demonstrates the power of ML in healthcare applications.